# Multivariate Methods

- Multivariate Data
- Missing Values
- Parameter Estimation
- Multivariate Normal Distribution
- Multivariate Classification
- Discrete Features
- Multivariate Regression



# Multivariate Data (1)

- The $l$ −dimensional feature vector
  - $\mathbf{x} = [x_1, \dots, x_l]^T$
- Mean vector
  - $\boldsymbol{\mu} = E\{\mathbf{x}\} = [\mu_1, \dots, \mu_l]^T$
    - $\mu_i = E\{x_i\}$
- Covariance
  - Between two random variables $X_i$ and $X_j$
    - $\sigma_{ij} = \mathrm{Cov}[X_i, X_j] = E\left\{\left(X_i - E(X_i)\right)\left(X_j - E(X_j)\right)\right\}$
      $= E\left(X_i X_j\right) - E(X_i)E(X_j)$
  - Measures the degree to which the two variables are related
    - In the range $[-\infty, \infty]$
    - In the same units as the features

# Multivariate Data (2)

- Uncorrelated
  - Two variables $X_i$ and $X_j$ are uncorrelated if their covariance is 0

- If two variables are independent
  - Their covariance is zero
    - $\because \sigma_{ij} = E\left\{(X_i - E(X_i))\left(X_j - E(X_j)\right)\right\}$

$$= \iint (X_i - E(X_i))\left(X_j - E(X_j)\right) p(X_i, X_j) dX_i dX_j$$

$$= \int (X_i - E(X_i)) p(X_i) dX_i \int \left(X_j - E(X_j)\right) p(X_j) dX_j = 0$$

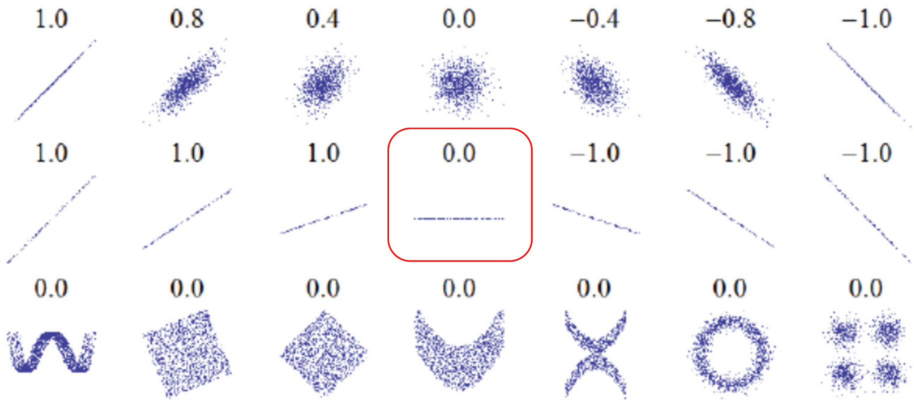
- But the converse is not true
  - Uncorrelated does NOT imply independent!!
    - $X_i$ and $X_j$ may be dependent even if $\sigma_{ij} = 0$



# Multivariate Data (3)

- Correlation
  - A normalized form of covariance
    - $\mathrm{Corr}[X_i, X_j] = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$
      - Ranges between $-1$ and $+1$
  - The measure responds only to linearity between features
    - One increases (or decreases), the other increases or decreases by a corresponding amount
    - If $X_j = aX_i + b, a > 0$
      - $\mathrm{Corr}[X_i, X_j] = \mathrm{Corr}[X_i, aX_i + b] = \frac{a\sigma_i^2}{\sigma_i \times a\sigma_i} = 1$
    - If $X_j = aX_i + b, a < 0$
      - $\mathrm{Corr}[X_i, X_j] = -1$
  - $\mathrm{Corr}[X_i, X_j]$ does NOT correspond to non-linear relationships between features

# Multivariate Data (4)



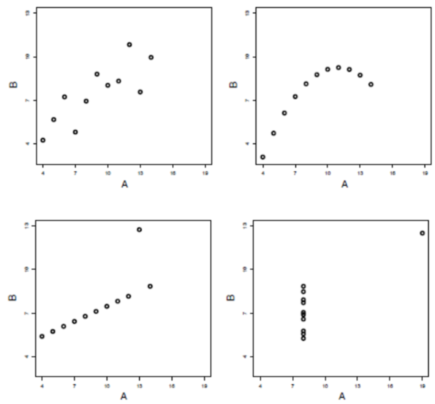



**Figure 2.12** Several sets of $(x, y)$ points, with the correlation coefficient of $x$ and $y$ for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of $Y$ is zero. Source: `http://en.wikipedia.org/wiki/File:Correlation_examples.png`

The 4 pairs of features all have the same correlation 0.816

Fig. 2.12 [Murphy]



# Multivariate Data (5)

- Covariance matrix
  - $\boldsymbol{\Sigma} = \mathrm{Cov}[\mathbf{x}] = E\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\} = E(\mathbf{x}\mathbf{x}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T$

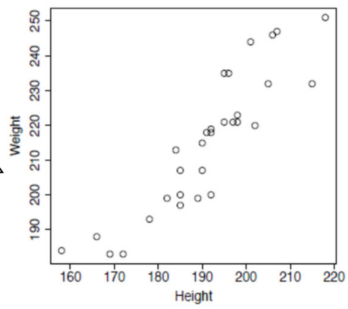
$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} \cdots & \sigma_{1l} \\ \sigma_{21} & \sigma_2^2 \cdots & \sigma_{2l} \\ \vdots & \ddots & \vdots \\ \sigma_{l1} & \sigma_{l2} \cdots & \sigma_l^2 \end{bmatrix}$$

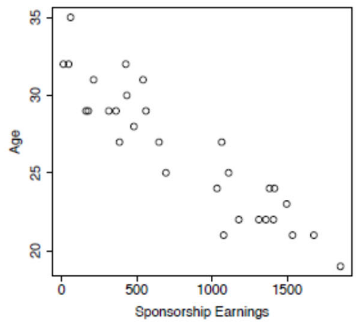
- Correlation matrix
  - $\mathrm{Corr}[\mathbf{x}] = \left(\mathrm{diag}(\boldsymbol{\Sigma})\right)^{-\frac{1}{2}} \boldsymbol{\Sigma} \left(\mathrm{diag}(\boldsymbol{\Sigma})\right)^{-\frac{1}{2}} = \begin{bmatrix} 1 & \rho_{12} \cdots & \rho_{1l} \\ \rho_{21} & 1 \cdots & \rho_{2l} \\ \vdots & \ddots & \vdots \\ \rho_{l1} & \rho_{l2} \cdots & 1 \end{bmatrix}$

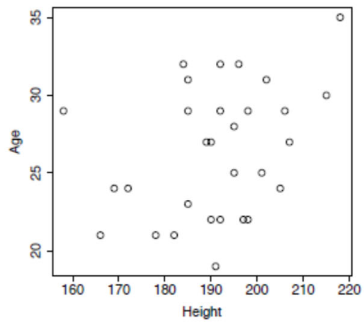
# Multivariate Data (6)

- Examples

Strong positive covariance




Strong negative covariance

Lack of strong covariance

$$cov(\text{HEIGHT}, \text{WEIGHT}) = \frac{7{,}009.9}{29} = 241.72$$

$$corr(Height, Weight) = \frac{241.72}{13.6 \times 19.8} = 0.898$$

corr(age,earnings) =-0.918

$$cov(\text{HEIGHT}, \text{AGE}) = \frac{570.8}{29} = 19.7$$

$$corr(Height, Age) = \frac{19.7}{13.6 \times 4.2} = 0.345$$

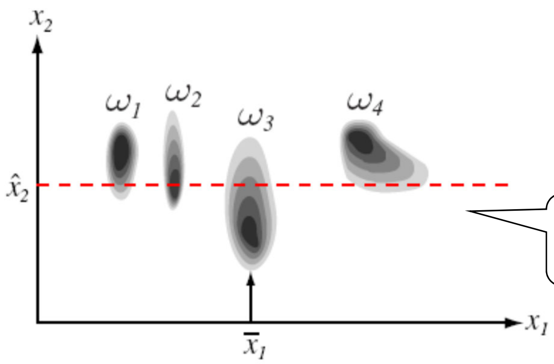


# Missing Values (1)

- Missing values
  - The number of available data is not the same for all features
    - Some samples have incomplete feature vectors
      - Partial responses in surveys of social sciences
      - In remote sensing, certain regions are covered by a subset of sensors
  - Omitting all incomplete feature vectors?
    - Not acceptable if there are many patterns with missing values
  - Completing the missing values (data imputation)?
    - By replacing with
      - Zeros
      - Class mean or median (mode, for discrete features) in the training set
      - Sample mean in the test set

# Missing Values (2)

- Example [Duda 01]
  - The feature $x_1$ is missing for a test pattern




If assuming the missing value is the mean of all $x_1$ features, we will classify the pattern as $\omega_3$

**FIGURE 2.22.** Four categories have equal priors and the class-conditional distributions shown. If a test point is presented in which one feature is missing (here, $x_1$) and the other is measured to have value $\hat{x}_2$ (red dashed line), we want our classifier to classify the pattern as category $\omega_2$, because $p(\hat{x}_2|\omega_2)$ is the largest of the four likelihoods. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Fig. 2.22 [Duda 01]



# Missing Values (3)

- Example 11.8 [Theodoridis 09, p. 615]
  - Consider the set with missing features
    - $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_5\} = \{\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1\\?\end{bmatrix}, \begin{bmatrix}0\\?\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}3\\1\end{bmatrix}\}$
  - If, substituting the missing values with the mean of the feature
    - $\mathbf{x}'_2 = \begin{bmatrix}1\\1\end{bmatrix}$ $\mathbf{x}'_3 = \begin{bmatrix}0\\1\end{bmatrix}$
  - If, measuring the distance using only the available features
    - The absolute distance
      - $d(\mathbf{x}_1, \mathbf{x}_2) = \frac{l}{l-(\#\text{missing features})} \sum_{\text{available features}} \text{distance} = \frac{2}{2-1} 1 = 2$
      - $d(\mathbf{x}_2, \mathbf{x}_3) = \frac{2}{2-1} 1 = 2$
      - $d(\mathbf{x}_1, \mathbf{x}_4) = \frac{2}{2-0} 4 = 4$

# Parameter Estimation Revisited

- Parametric model $p(\mathbf{x}|C_i) \equiv p(\mathbf{x}|C_i; \boldsymbol{\theta}_i)$
  - Maximum-likelihood estimation (MLE)
    - Maximizing the probability of obtaining the samples $X$ observed
    - $\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{argmax}\, p(X|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{argmax} \prod_{k=1}^{N} p(\mathbf{x}_k|\boldsymbol{\theta})$
      - $L(\boldsymbol{\theta}) \equiv ln\, p(X|\boldsymbol{\theta}) = \sum_{k=1}^{N} \ln p(\mathbf{x}_k|\boldsymbol{\theta})$
      - Let $\nabla_{\boldsymbol{\theta}} L \equiv \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$
  - Maximum A Posteriori (MAP) estimation
    - $\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{argmax}\, p(\boldsymbol{\theta}|X) = \underset{\boldsymbol{\theta}}{argmax}\, p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})$



# Multivariate Normal Distribution (1)

- $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \mathbf{x} = [x_1, \ldots, x_l]^T$
  - $p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{l}{2}}\sqrt{|\boldsymbol{\Sigma}|}} \exp(-\frac{(\mathbf{x}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2})$

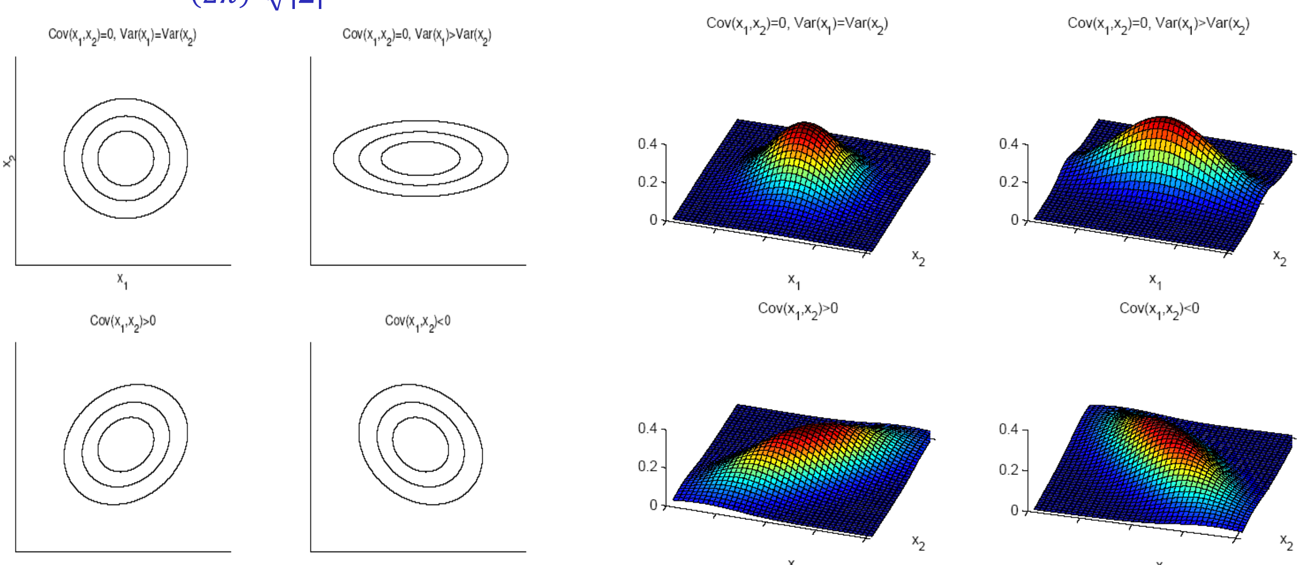



Fig. 5.2 [Alpaydin, 2020]

# Multivariate Normal Distribution (2)

- The Gaussian Case 1: unknown $\boldsymbol{\mu}$
  - Suppose the samples are drawn from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
    - $L(\boldsymbol{\mu}) = \sum_{i=1}^{N} \ln p(\mathbf{x}_i|\boldsymbol{\mu})$
      $= \sum_{i=1}^{N} \left\{-\frac{1}{2}\ln\left((2\pi)^l|\Sigma|\right) - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}$
      $= -\frac{Nl}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \sum_{i=1}^{N}\left\{\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}$
  - From
    - $\frac{\partial}{\partial\boldsymbol{\mu}}\{(\mathbf{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\} = \frac{\partial}{\partial\mathbf{y}_i}\{\mathbf{y}_i^T\Sigma^{-1}\mathbf{y}_i\}\frac{\partial\mathbf{y}_i}{\partial\boldsymbol{\mu}} = -\left(\Sigma^{-1} + \Sigma^{-T}\right)\mathbf{y}_i$
      $= -2\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$

Let $\mathbf{y}_i = \mathbf{x}_i - \boldsymbol{\mu}$

$\frac{\partial}{\partial\mathbf{x}}[\boldsymbol{x}^T\boldsymbol{M}\boldsymbol{x}] = [\boldsymbol{M} + \boldsymbol{M}^T]\boldsymbol{x}$

  - We have
    - $\nabla_{\boldsymbol{\mu}}L = -\frac{1}{2}\sum_{i=1}^{N}\{-2\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\} = \Sigma^{-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{0}$
  - ML estimate
    - $\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$



# Multivariate Normal Distribution (3)

- The Gaussian Case 1: unknown $\boldsymbol{\mu}$
  - Suppose the unknown $\boldsymbol{\mu}$ is known to be normally distributed
    - $p(\boldsymbol{\theta}) = p(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$
  - The posterior probability
    - $p(\boldsymbol{\theta}|X) = p(\boldsymbol{\mu}|X) = \cdots = N(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$
      - $\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_0\left(\boldsymbol{\Sigma}_0 + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\mu}}_{ML} + \frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_0 + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}_0$
      - $\boldsymbol{\Sigma}_N = \boldsymbol{\Sigma}_0\left(\boldsymbol{\Sigma}_0 + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\frac{1}{N}\boldsymbol{\Sigma}$
  - MAP estimation

A linear combination of ML mean and the prior mean $\mu_0$

    - $\hat{\boldsymbol{\mu}}_{MAP} = \boldsymbol{\mu}_N = \boldsymbol{\Sigma}_0\left(\boldsymbol{\Sigma}_0 + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\mu}}_{ML} + \frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_0 + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}_0$
  - The Bayes' estimation
    - $p(\mathbf{x}|X) = \int p(\mathbf{x}|\boldsymbol{\mu})p(\boldsymbol{\mu}|X)d\boldsymbol{\mu} = \cdots = N(\boldsymbol{\mu}_N, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_N)$

The increased variance results from our lack of exact knowledge of $\boldsymbol{\mu}$

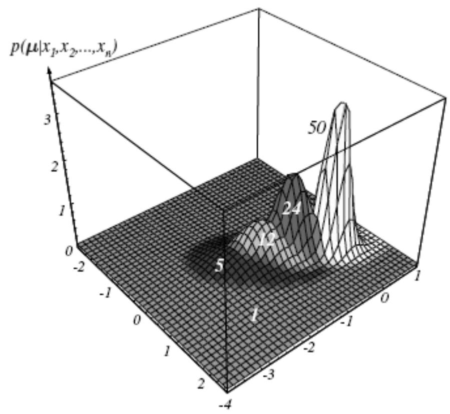
# Multivariate Normal Distribution (4)

- The Gaussian Case 1: unknown $\boldsymbol{\mu}$

The posterior $p(\boldsymbol{\mu}|X) = N(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$ with different numbers of training samples

$$p(\mathbf{x}_i|\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\mu} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.1 \end{bmatrix}$$

$p(\boldsymbol{\mu}) \sim N(\mathbf{0}, 0.1\mathbf{I})$ $\quad$ $p(\boldsymbol{\mu}|X)$



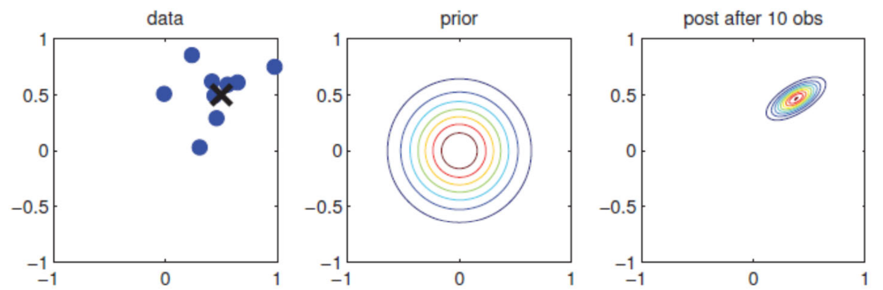




Figure 4.13 Illustration of Bayesian inference for the mean of a 2d Gaussian. (a) The data is generated from $\mathbf{y}_i \sim \mathcal{N}(\mathbf{x}, \boldsymbol{\Sigma}_y)$, where $\mathbf{x} = [0.5, 0.5]^T$ and $\boldsymbol{\Sigma}_y = 0.1[2, 1; 1, 1]$). We assume the sensor noise covariance $\boldsymbol{\Sigma}_y$ is known but $\mathbf{x}$ is unknown. The black cross represents $\mathbf{x}$. (b) The prior is $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, 0.1\mathbf{I}_2)$. (c) We show the posterior after 10 data points have been observed. Figure generated by gaussInferParamsMean2d.

Fig. 4.13 [Murphy 2012]



# Multivariate Normal Distribution (5)

- The Gaussian Case 2: unknown $\boldsymbol{\mu}$ and $\Sigma$
  - $L(\boldsymbol{\mu}, \Sigma) = \sum_{i=1}^{N} \ln p(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma)$
    $= -\frac{Nl}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \sum_{i=1}^{N}\left\{\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}$
  - Let $\nabla_{\boldsymbol{\mu}} L = 0$
    - $\Longrightarrow \hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$
  - Let $\nabla_{\Sigma} L = 0$
    - Rewrite the log-likelihood term (let $\Lambda = \Sigma^{-1}$)
      - $L = const + \frac{N}{2}\ln|\Lambda| - \frac{1}{2}\sum_{i=1}^{N} tr\{\Lambda(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{x}_i - \boldsymbol{\mu})\}$
      - $\nabla_{\Lambda} L = \frac{N}{2}\Lambda^{-T} - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T = \mathbf{0}$
      - $\Lambda^{-T} = \Lambda^{-1} = \Sigma = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$
    - $\Longrightarrow \hat{\Sigma}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$

$$tr(c) = c$$
$$tr(\mathbf{AB}) = tr(\mathbf{BA})$$
$$\mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x} = tr(\mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x}) = tr(\mathbf{A}\mathbf{x}\mathbf{x}^{\mathbf{T}})$$
$$\frac{\partial}{\partial \mathbf{X}}\ln|\mathbf{X}| = (\mathbf{X}^{-1})^T$$
$$\frac{\partial}{\partial \mathbf{X}} tr(\mathbf{X}^{\mathbf{T}}\mathbf{A}) = \mathbf{A}$$

# Multivariate Normal Distribution (6)

- The Gaussian Case 2: unknown $\boldsymbol{\mu}$ and $\Sigma$
  - The full covariance matrix is singular if $N < l$
    - $\hat{\Sigma}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\boldsymbol{T}}$
  - Strategies for preventing overfitting
    - Use a diagonal covariance matrix for each class
      - Features are assumed conditionally independent
      - Naïve Bayes classifier
    - Force the full covariance matrix to be the same for all classes
      - Linear discriminant analysis (i.e., Case 3 in Ch3)
    - Project the data into a low dimensional subspace and fit the Gaussians there



# Multivariate Classification (1)

- Assume $p(\mathbf{x}|C_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
  - Given the training data set $\{X_1, X_2, \dots, X_K\}$
  - $g_i(\mathbf{x}) = ln\, p(\mathbf{x}|C_i) + \ln P(C_i)$

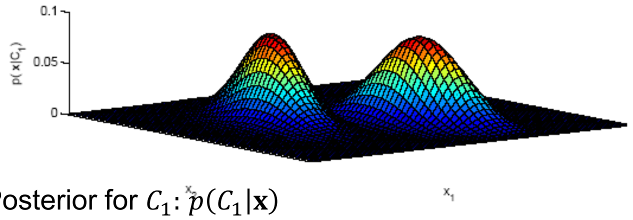
$$= -\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^{\boldsymbol{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2} - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(C_i)$$

  - We estimate the unknown parameters for each class separately
    - $\hat{\boldsymbol{\mu}}_i = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j$
    - $\hat{\Sigma}_i = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\boldsymbol{T}}$
    - $\hat{P}(C_i) = \frac{N_i}{N_1 + \cdots + N_K}$
  - The discriminant function becomes
    - $g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\hat{\boldsymbol{\mu}}_i)^{\boldsymbol{T}}\hat{\Sigma}_i^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_i)}{2} - \frac{1}{2}\ln|\hat{\Sigma}_i| + \ln\hat{P}(C_i)$

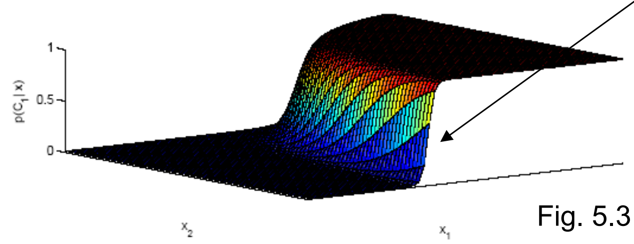
# Multivariate Classification (2)

- Review of Ch3
  - Case 4 (quadratic discriminant): $\mathbf{\Sigma}_i = \text{arbitrary}$
    - $g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\hat{\boldsymbol{\mu}}_i)^T\hat{\Sigma}_i^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_i)}{2} - \frac{1}{2}\ln\left|\hat{\Sigma}_i\right| + \ln\hat{P}(C_i) = \mathbf{x}^T\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^T\mathbf{x} + w_{i0}$

Likelihood densities $p(\mathbf{x}|C_i)$



Posterior for $C_1$: $p(C_1|\mathbf{x})$



discriminant: $p(C_1|\mathbf{x}) = 0.5$

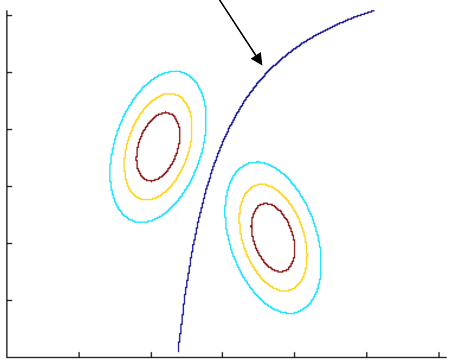
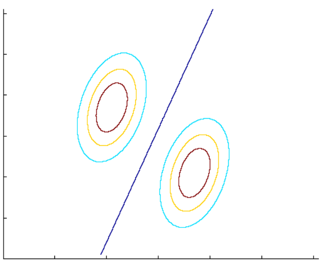


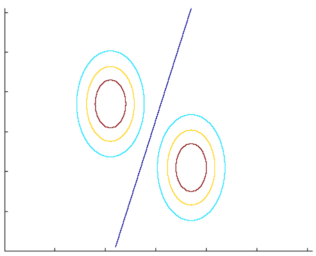
Fig. 5.3 [Alpaydin, 2020]

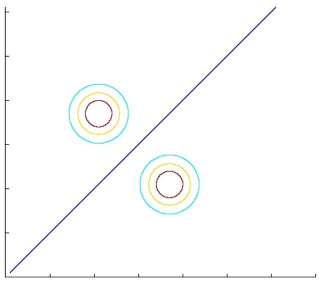


# Multivariate Classification (3)

- Review of Ch3
  - Case 3 (linear discriminant): $\mathbf{\Sigma}_i = \mathbf{\Sigma}$
    - $g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\hat{\boldsymbol{\mu}}_i)^T\hat{\Sigma}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_i)}{2} + \ln\hat{P}(C_i) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$
  - Case 2 (linear discriminant): $\mathbf{\Sigma}_i = \mathbf{\Sigma} = diag\left(\sigma_1^2, \sigma_2^2, \dots, \sigma_l^2\right)$
    - $g_i(\mathbf{x}) = -\frac{1}{2}\sum_{j=1}^{l}\left(\frac{x_j - \hat{\mu}_{i,j}}{\hat{\sigma}_j}\right)^2 + \ln\hat{P}(C_i) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$
  - Case 1 (linear discriminant): $\mathbf{\Sigma}_i = \mathbf{\Sigma} = \sigma^2\mathbf{I}$
    - $g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\hat{\boldsymbol{\mu}}_i\|^2}{2\hat{\sigma}^2} + \ln\hat{P}(C_i) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$







Figs. 5.4-5.6 [Alpaydin, 2020]

# Multivariate Classification (4)

- Bayesian classification for normal distribution

| *Assumption* | *Covariance matrix* | *#parameters* |
|---|---|---|
| Shared, Hyperspheric (case 1) | $\mathbf{\Sigma}_i = \mathbf{\Sigma} = \sigma^2\mathbf{I}$ | $1$ |
| Shared, Axis-aligned (case 2) | $\mathbf{\Sigma}_i = \mathbf{\Sigma}$ with $\sigma_{ij} = 0$ | $l$ |
| Shared, Hyperellipsoidal (case 3) | $\mathbf{\Sigma}_i = \mathbf{\Sigma}$ | $l(l+1)/2$ |
| Different, Hyperellipsoidal (case 4) | $\mathbf{\Sigma}_i$ | $Kl(l+1)/2$ |

Table 5.1 [Alpaydin, 2020]

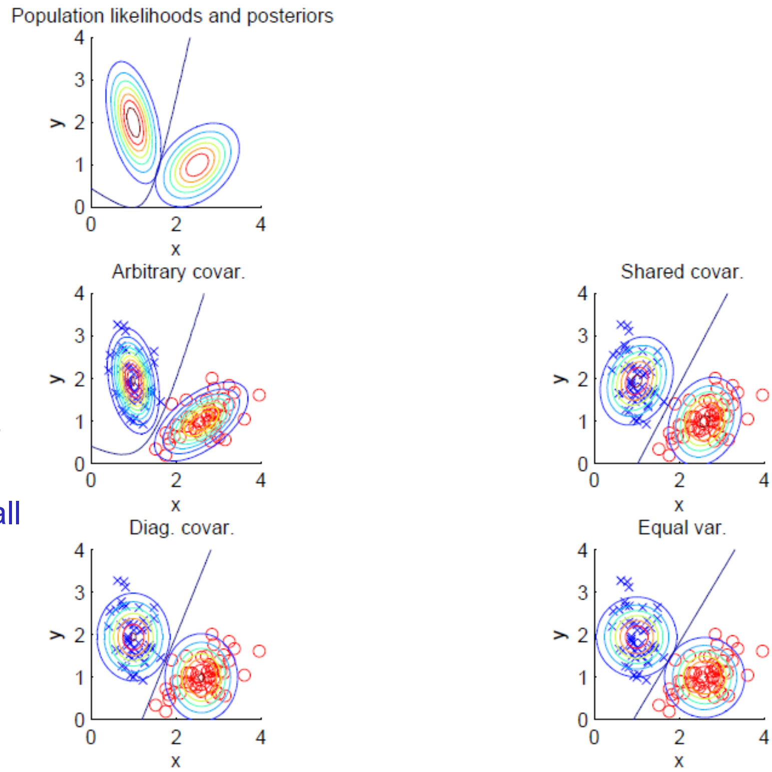
- Bias/variance dilemma
  - When increasing complexity (less restricted $\mathbf{\Sigma}$)
    - Bias ↓
    - Variance ↑
  - When assuming simple models
    - Bias ↑
    - Variance ↓



# Multivariate Classification (5)

- Tuning complexity
  - Depends on
    - The data at hand
    - The amount of data
  - Small dataset
    - Even if $\mathbf{\Sigma}_i$ are different
    - Better assume a shared $\mathbf{\Sigma}$
      - Fewer parameters to be estimated from data of all classes

Fig. 5.7 [Alpaydin, 2020]



# Multivariate Classification (6)

- Regularized discriminant analysis (RDA)
  - A weighted average of three special cases (cases 1, 3, and 4)
    - $\hat{\mathbf{\Sigma}}'_i = \alpha\sigma^2\mathbf{I} + \beta\hat{\mathbf{\Sigma}} + (1-\alpha-\beta)\hat{\mathbf{\Sigma}}_i$
    - $\alpha$: a shrinkage parameter
      - Covariance matrix updates
    - $\beta$: a complexity parameter
      - An intermediate between linear and quadratic discriminant
  - $\alpha, \beta$ are chosen by cross-validation
    - When $\alpha = \beta = 0$
      - (case 4) quadratic classifier
    - When $\alpha = 0,\ \beta = 1$
      - (case 3) linear classifier
    - When $\alpha = 1,\ \beta = 0$
      - (case 1) linear classifier



# Discrete Features (1)

- Discrete features – binary case
  - The feature vector $\mathbf{x} = [x_1, \dots, x_l]^T$ and its indicator $\mathbf{y} = [y_1, \dots, y_K]^T$
    - Each $x_j \in \{0,1\}$ is a Bernoulli random variable with
      - $p_{ij} \equiv p(x_j = 1|C_i), \quad y_i = \begin{cases} 1, \mathbf{x} \in C_i \\ 0, \mathbf{x} \notin C_i \end{cases}$
    - $p(\mathbf{x}|C_i) = p(x_1, x_2, \dots, x_l|C_i) = \prod_{j=1}^{l} p_{ij}^{x_j}(1-p_{ij})^{(1-x_j)}$
  - Given an iid sample $X = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$
    - The ML estimate (Ch4)
      - $\hat{p}_{ij} = \frac{\sum_m x_{m,j} y_{m,i}}{\sum_m y_{m,i}}$
  - The discriminant function is linear
    - $g_i(\mathbf{x}) = \ln P(\mathbf{x}|C_i) + \ln P(C_i)$

$$= \sum_{j=1}^{l}[x_j \ln \hat{p}_{ij} + (1-x_j)\ln(1-\hat{p}_{ij})] + \ln P(C_i)$$

$$= \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}$$

# Discrete Features (2)

- Example 2.10 (p.60 [Theodoridis 09])
  - Discrete binary feature & two-category case
    - The feature vector $\mathbf{x} = [x_1, \dots, x_l]^T$ with binary attributes $x_j \in \{0,1\}$
      - $p_{1j} \equiv p(x_j = 1|C_1)$ and $p_{2j} \equiv p(x_j = 1|C_2)$
    - Adopting Naïve Bayesian assumption (i.e., conditional independent)
      - $p(\mathbf{x}|C_i) = \prod_{j=1}^{l} p_{ij}^{x_j}(1-p_{ij})^{(1-x_j)}, i = 1,2$
      - The number of required estimates is $2l$ (i.e. $\hat{p}_{1j}$ and $\hat{p}_{2j}, j = 1, \dots, l$)
    - The discriminant function
      - $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) = \sum_{j=1}^{l}[x_j \ln\frac{\hat{p}_{1j}}{\hat{p}_{2j}} + (1-x_j)\ln\frac{1-\hat{p}_{1j}}{1-\hat{p}_{2j}}] + \ln\frac{P(C_1)}{P(C_2)}$

        $= \mathbf{w}^T\mathbf{x} + w_0$
        - » $\mathbf{w} = \left[\ln\frac{\hat{p}_{11}(1-\hat{p}_{11})}{\hat{p}_{21}(1-\hat{p}_{21})}, \dots, \frac{\hat{p}_{1l}(1-\hat{p}_{1l})}{\hat{p}_{2l}(1-\hat{p}_{2l})}\right]^T$
        - » $w_0 = \sum_{j=1}^{l}[\ln\frac{1-\hat{p}_{1j}}{1-\hat{p}_{2j}})] + \ln\frac{P(C_1)}{P(C_2)}$

If $p_{1j} = p_{2j}$, then $w_j = 0$
$\Longrightarrow x_j$ gives no information

If $p_{1j} > p_{2j}$, then $w_j > 0$
$\Longrightarrow x_j$ contributes votes to $C_1$



# Discrete Features (3)

- Discrete features – general case
  - The feature vector $\boldsymbol{x} = [x_1, \dots, x_l]^T$and its indicator $\mathbf{y} = [y_1, \dots, y_K]^T$
    - Each $x_j \in \{v_1, \dots, v_{n_j}\}$ has $n_j$ states
    - Define 0/1 dummy variables as
      - $z_{jk} \equiv \begin{cases} 1, \text{if } x_j = v_k \\ 0, \text{otherwise} \end{cases}$ and $\sum_{k=1}^{n_j} z_{jk} = 1$
    - Let $p_{ijk} \equiv P(z_{jk} = 1|C_i) = P(x_j = v_k|C_i)$
    - $p(\mathbf{x}|C_i) = p(x_1, x_2, \dots, x_l|C_i) = \prod_{j=1}^{l}\prod_{k=1}^{n_j} p_{ijk}^{z_{jk}}$
    - The ML estimate (Ch4)
      - $\hat{p}_{ijk} = \frac{\sum_m z_{m,jk} y_{m,i}}{\sum_m y_{m,i}}$
  - The discriminant function is
    - $g_i(\mathbf{x}) = \sum_{j=1}^{l}\sum_k [z_{jk}\ln\hat{p}_{ijk}] + \ln P(C_i)$

# Multivariate Regression (1)

- Multivariate regression
  - $X = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$, $\mathbf{x}_i \in R^l$
    - $\mathbf{y} = f(\mathbf{x}) + \varepsilon, \varepsilon \sim N(0, \sigma^2)$
  - To approximate the unknown $f(\mathbf{x})$ by the estimator $g(\mathbf{x}|\boldsymbol{\theta})$
    - $p(y|\mathbf{x}, \boldsymbol{\theta}) \sim N(y|g(\mathbf{x}|\boldsymbol{\theta}), \sigma^2)$
    - $L(\boldsymbol{\theta}) \equiv \ln p(X|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$

$$= \sum_{i=1}^{N} \ln\left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{\left(y_i - g(\mathbf{x}_i|\boldsymbol{\theta})\right)^2}{2\sigma^2}\right)\right)$$

$$= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - g(\mathbf{x}_i|\boldsymbol{\theta})\right)^2$$

  - Maximizing $L(\boldsymbol{\theta})$ = minimizing the sum of squared error
    - $E(\boldsymbol{\theta}|X) = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - g(\mathbf{x}_i|\boldsymbol{\theta})\right)^2$



# Multivariate Regression (2)

- Multivariate linear regression
  - Assuming that $g(\mathbf{x}|\boldsymbol{\theta})$ is linear
    - $g(\mathbf{x}|w_0, w_1, \dots, w_l) = w_0 + w_1 x_1 + \dots + w_l x_l = \mathbf{w}^T \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}$
      - $\mathbf{x} = [x_1, x_2, \dots, x_l]$
  - The sum of squared error
    - $E(w_0, w_1, \dots, w_l|X) = \frac{1}{2} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_1 + \dots - w_l x_l)^2$
  - Let
    - $\boldsymbol{X} = \begin{bmatrix} 1 & \mathbf{x}_1^T \\ & \vdots \\ 1 & \mathbf{x}_N^T \end{bmatrix}$ $\quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ $\quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_l \end{bmatrix}$
    - $\frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} = -\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0$
      - $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$ (normal equation)
      - $\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

Same as in polynomial regression if we define $\mathbf{x} = [x, x^2, \dots, x^l]$

We can define any nonlinear function using basis functions, e.g., $\mathbf{x} = [x, \sin(x), \exp(x^2)]$